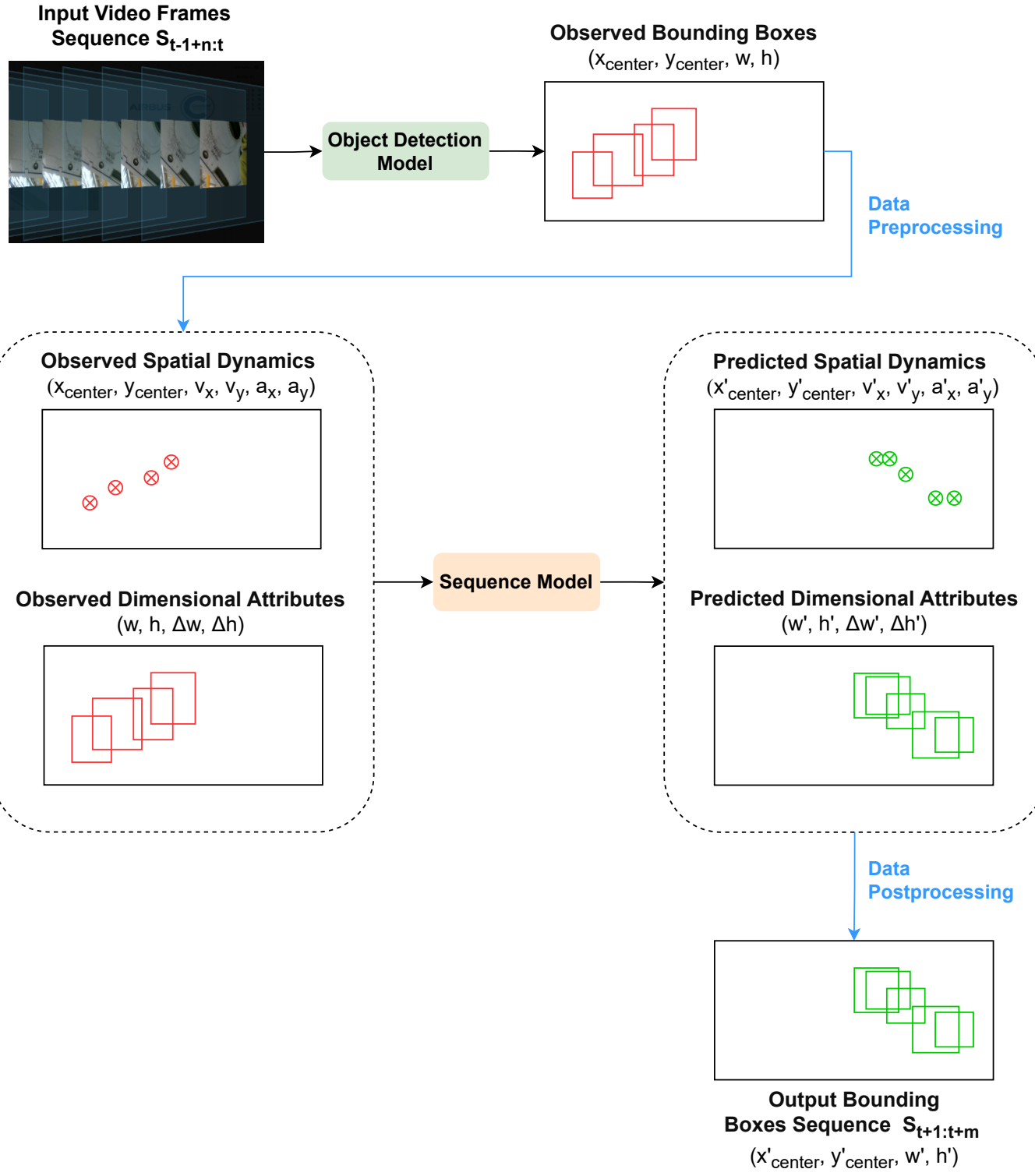
Input Video Frames Sequence $S_{t-1+n:t}$
Object Detection Model
Observed Bounding Boxes
(x_{center}, y_{center}, w, h)
Data Preprocessing
Observed Spatial Dynamics
(x_{center}, y_{center}, v_x, v_y, a_x, a_y)
Observed Dimensional Attributes
(w, h, Δw, Δh)
Sequence Model
Predicted Spatial Dynamics
(x'_{center}, y'_{center}, v'_x, v'_y, a'_x, a'_y)
Predicted Dimensional Attributes
(w', h', Δw', Δh')
Data Postprocessing
Output Bounding Boxes Sequence $S_{t+1:t+m}$
(x'_{center}, y'_{center}, w', h')